IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Neil
Oct 8

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
- - Data collection using API
- - Data wrangling
- - Data visualization for exploration
- -Data analysis with SQL and Python
- - Model building with ML

Summary of all results

- Results of launch success over time

- Model prediction perform well

Introduction

- Project background and context

-SpaceX's reusable rocket stages, especially the Falcon 9, have drastically cut the cost of space missions by enabling multiple launches with the same rocket stage. However, accurately predicting when a stage can be reused without costly refurbishment remains a challenge due to various mission-specific factors.

- Problems we want to solve

  -Prediction of Reusability

  -Launch Cost Optimization

  -Effect of Mission-specific Factors

Section 1

Methodology

Methodology

Executive Summary

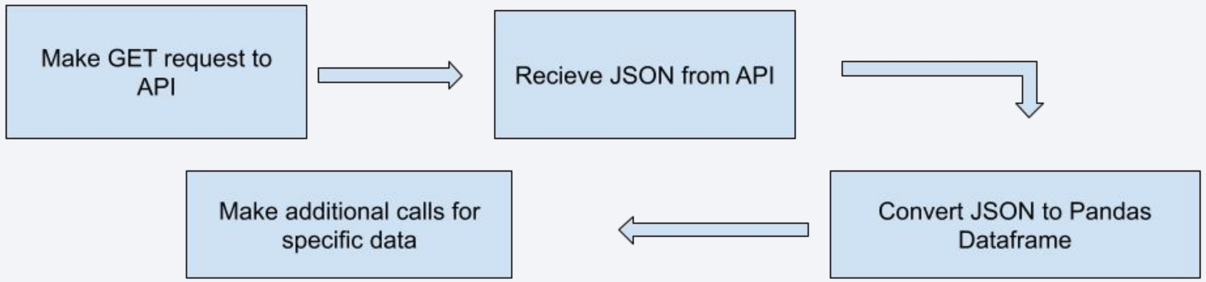
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

In this project, data was collected from the SpaceX REST API and additional web scraping. The API provided detailed information on rocket launches, including the rocket used, payload delivered, launch and landing specifications, and landing outcomes. The endpoint `api.spacexdata.com/v4/launches/past` was used to gather data on past launches, which was then processed into a structured format using Python libraries such as `requests` and `json_normalize`. Additionally, data was scraped from Wiki pages using BeautifulSoup for further analysis. Issues like null values in PayloadMass were handled by calculating and replacing missing data with the mean.

Data Collection – SpaceX API

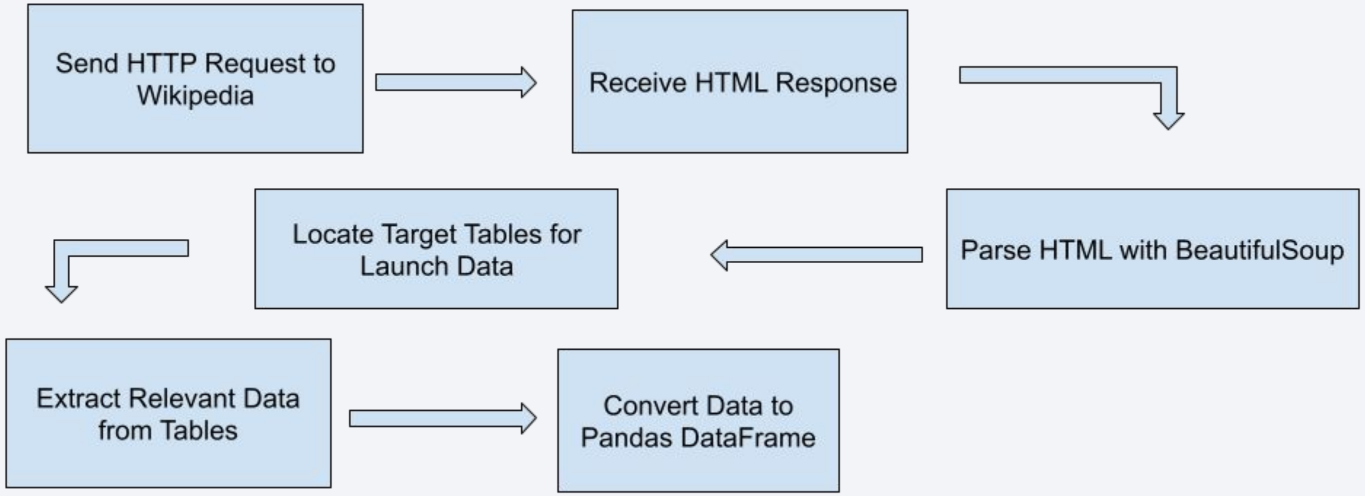
- Data was collected from the SpaceX REST API by targeting the /v4/launches/past endpoint. A GET request was used to retrieve JSON data on past launches. This data was then converted into a Pandas DataFrame using the json_normalize function for further analysis. Additional API calls were made to gather specific details, such as rocket and payload information.



- https://github.com/Barracuda108/Data-Science-Capstone/blob/293a7ea5888c1b0c3df1fea2453802958520b82b/1-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Web scraping was conducted using the BeautifulSoup library to extract Falcon 9 launch data from relevant Wikipedia tables. The HTML content was parsed to locate and retrieve valuable launch records. This data was then organized and converted into a Pandas DataFrame for further analysis. Web scraping complemented the API data collection by providing additional context and details on Falcon 9 launches.

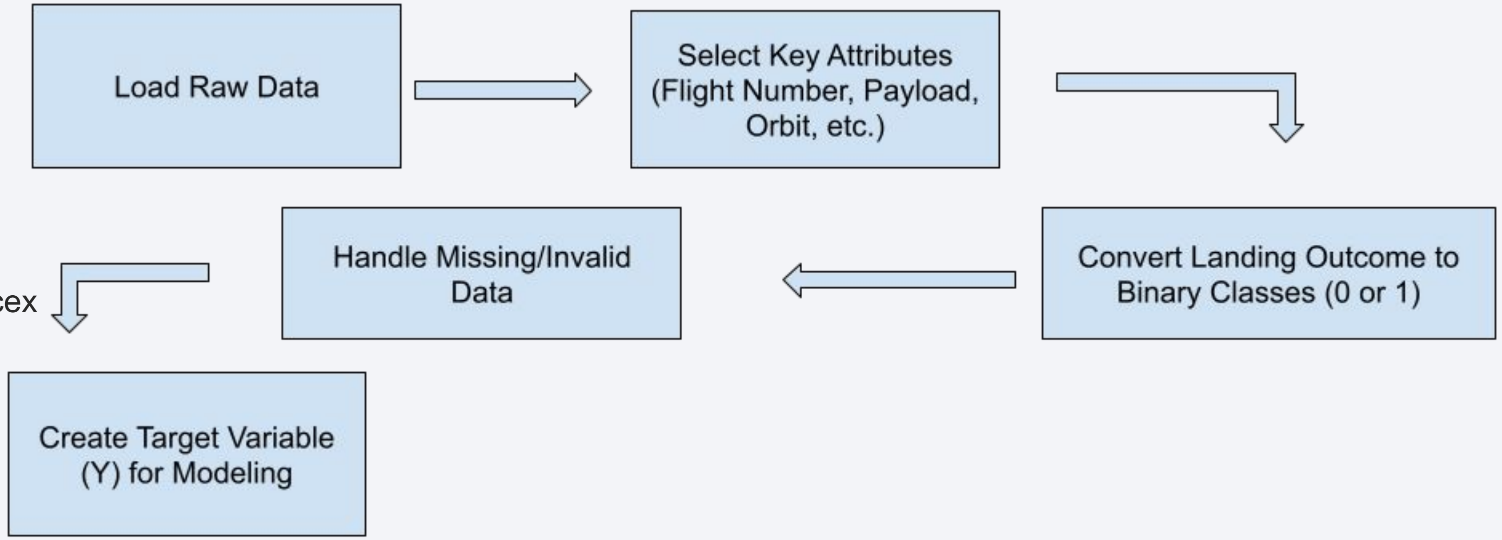


- https://github.com/Barracuda108/Data-Science-Capstone/blob/293a7ea5888c1b0c3df1fea2453802958520b82b/2-jupyter-labs-webscraping.ipynb

Data Wrangling

Key attributes like Flight Number, Booster Version, Payload Mass, Orbit, and Launch Site were processed. Landing Outcomes were converted into binary classes: 1 for successful landings and 0 for failures, creating the target variable (Y) for modeling.

https://github.com/Barracuda108/Data-Science-Capstone/blob/293a7ea5888c1b0c3df1fea2453802958520b82b/3-jupyter-labs-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

Visualize the launch success yearly trend with a line chart
FlightNumber vs. PayloadMass - scatter plot
Catplot to plot FlightNumber vs LaunchSite - category plot
Launch sites and their payload mass- category plot
Relationship between success rate and orbit type - bar plot
FlightNumber and Orbit type - category plot
Payload Mass and Orbit - category plot

https://github.com/Barracuda108/Data-Science-Capstone/blob/293a7ea5888c1b0c3df1fea245380295852 0b82b/5-edadataviz.ipynb

EDA with SQL

- **Display the names of the unique launch sites in the space mission**
- **Display 5 records where launch sites begin with the string 'CCA**
- **Display the total payload mass carried by boosters launched by NASA (CRS)**
- **Display average payload mass carried by booster version F9 v1.1**
- **List the date when the first succesful landing outcome in ground pad was acheived.**
- **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
- **List the total number of successful and failure mission outcomes**
- **List the names of the booster_versions which have carried the maximum payload mass. Use a subquery**
- **List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.**
- **Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.**

https://github.com/Barracuda108/Data-Science-Capstone/blob/293a7ea5888c1b0c3df1fea2453802958520b82b/4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Mark all launch sites on a map using markers

Mark the success/failed launches for each site on the map using circles and markers

Calculate the distances between a launch site to its proximities using lines

These objects were added to the map to visually analyze launch site locations, outcomes, and proximities, providing insights into patterns that can help identify optimal launch sites.

https://github.com/Barracuda108/Data-Science-Capstone/blob/293a7ea5888c1b0c3df1fea2453802958520b82b/6-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

1. **Dropdown for Launch Site Selection**:
   - Added a **dropdown menu** to select from different launch sites or view all sites.
   - **Why**: To allow users to filter data by launch site and view specific success rates for each site.
2. **Success Rate Pie Chart**:
   - A **pie chart** that shows the number of successful and failed launches.
   - **Why**: To give a quick visual representation of the success rate for selected launch sites.
3. **Range Slider for Payload Mass**:
   - A **range slider** for selecting a payload mass range (0 to 10,000 kg).
   - **Why**: To explore the correlation between payload mass and launch success, letting users filter based on payload.
4. **Scatter Plot for Payload vs. Success**:
   - A **scatter plot** that shows the relationship between payload mass and launch success, with **color-coded booster versions**.
   - **Why**: To visually identify how payload mass and booster version affect launch outcomes, helping in deeper analysis.

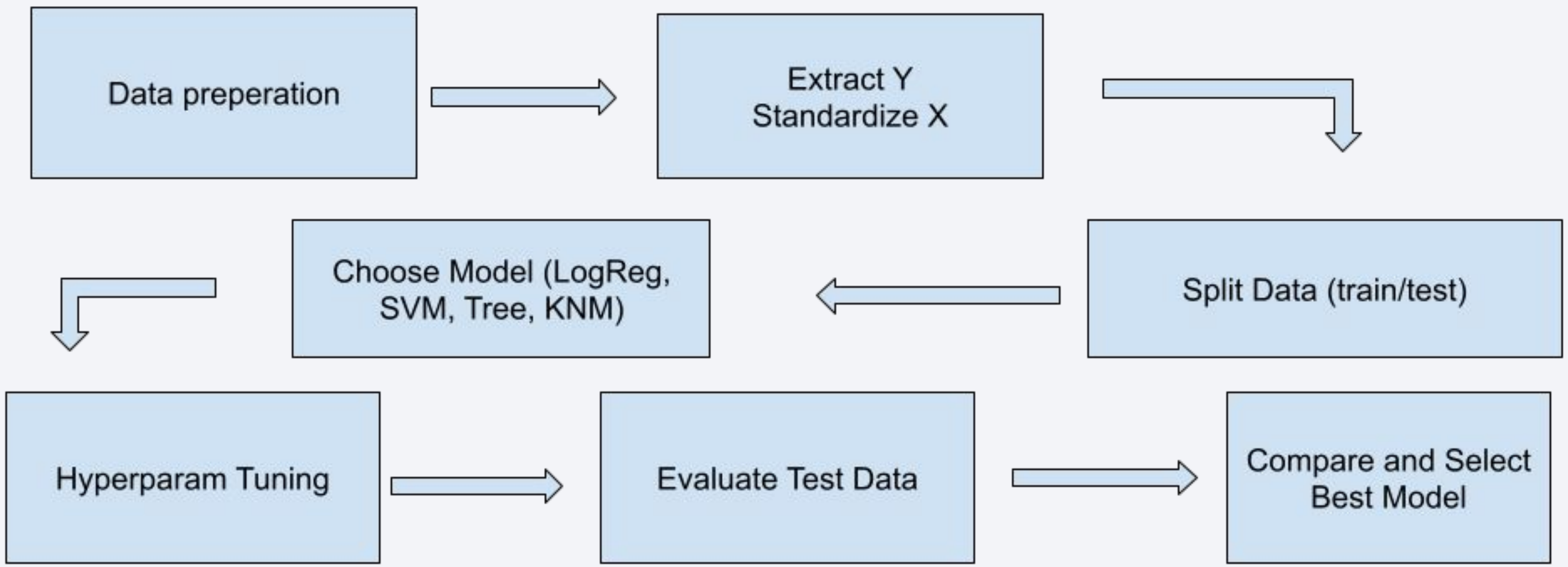
These plots and interactions were added to allow users to explore relationships between launch site, payload mass, and success rate, facilitating data-driven insights into SpaceX launches.

https://github.com/Barracuda108/Data-Science-Capstone/blob/293a7ea5888c1b0c3df1fea2453802958520b82b/7-spacex_dash_app.py

Predictive Analysis (Classification)

The model development process involved creating the target variable, standardizing the data, and splitting it into training and testing sets. Logistic Regression, SVM, Decision Trees, and KNN models were trained and optimized using GridSearchCV with 10-fold cross-validation. Each model was evaluated for accuracy on the test data, and the best-performing model was selected based on the highest accuracy and lowest error rate.

https://github.com/Barracuda108/Data-Science-Capstone/blob/293a7ea5888c1b0c3df1fea2453802958520b82b/8-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

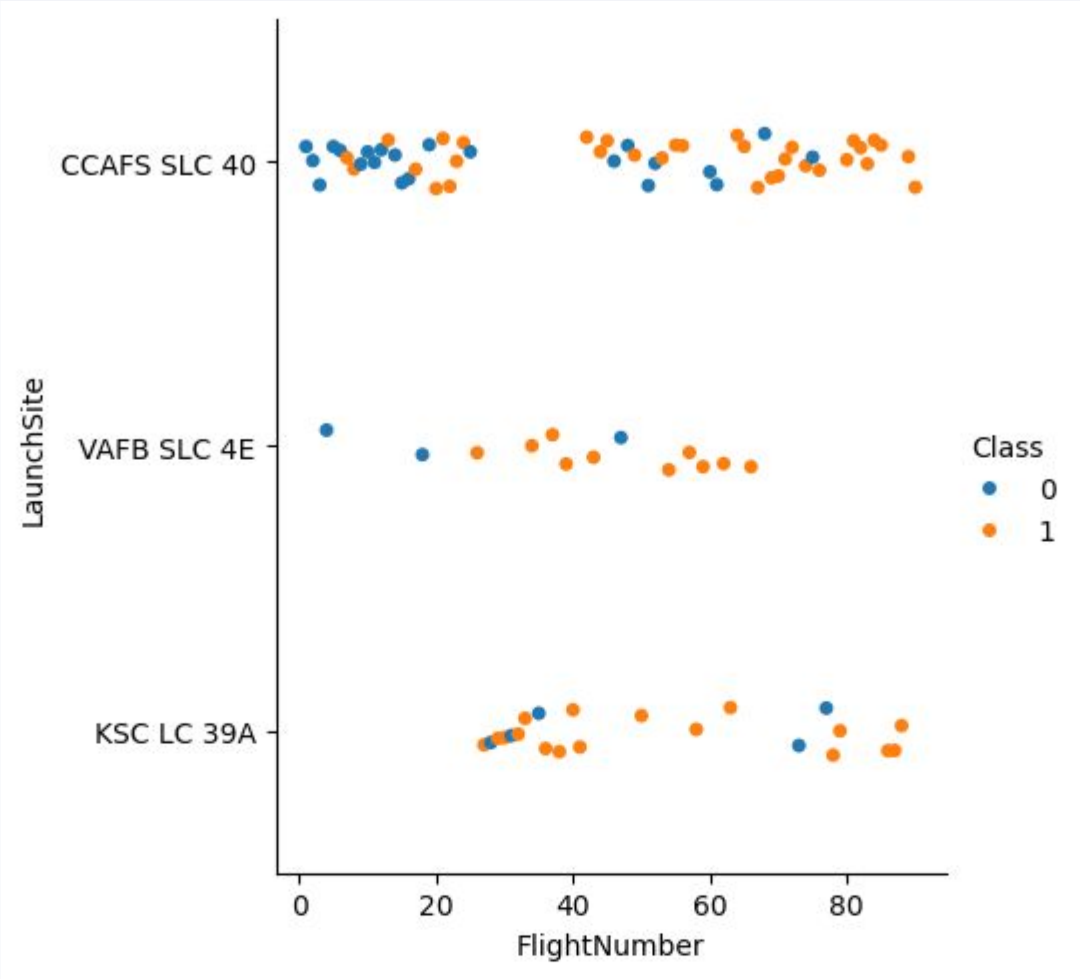
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

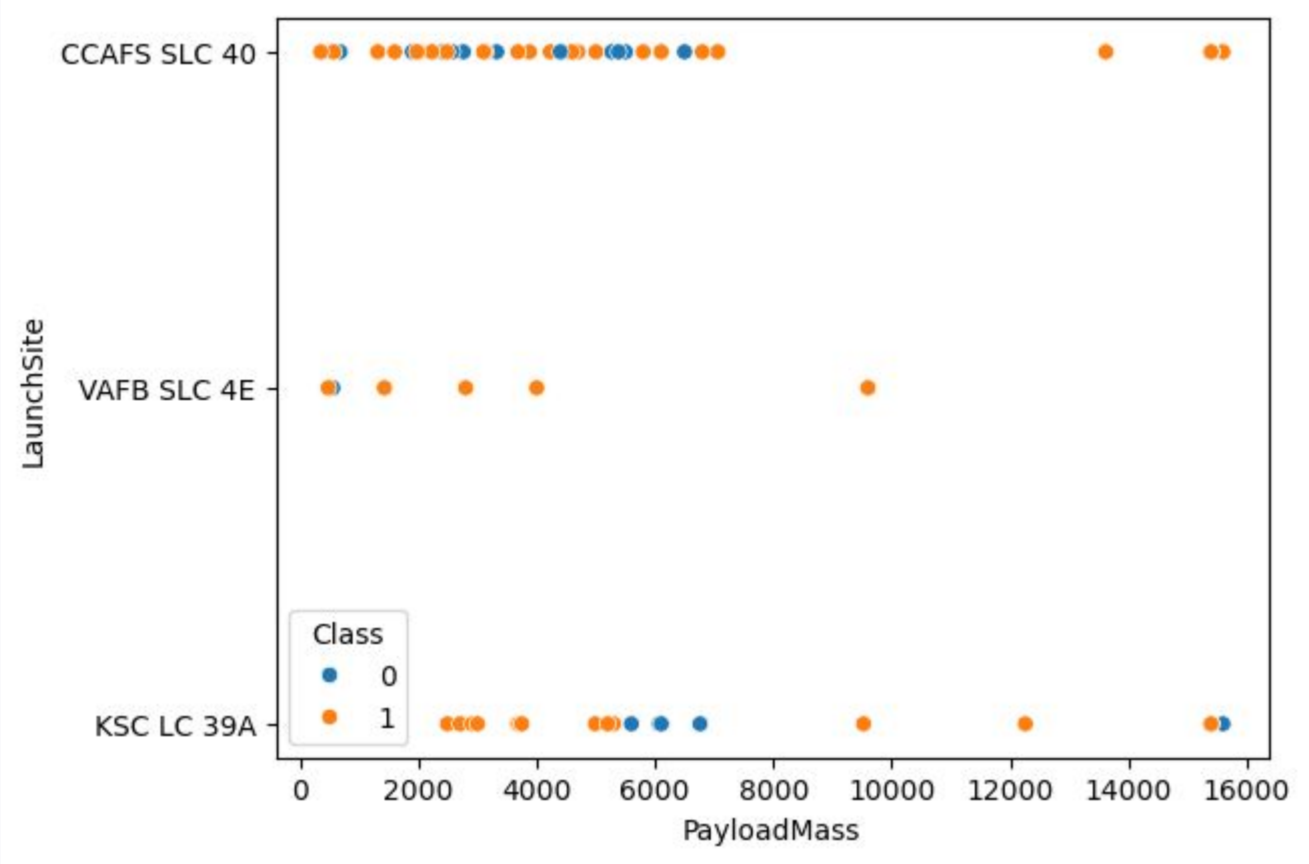
Flight Number vs. Launch Site

- Show the screenshot of the scatter plot with explanations



Payload vs. Launch Site

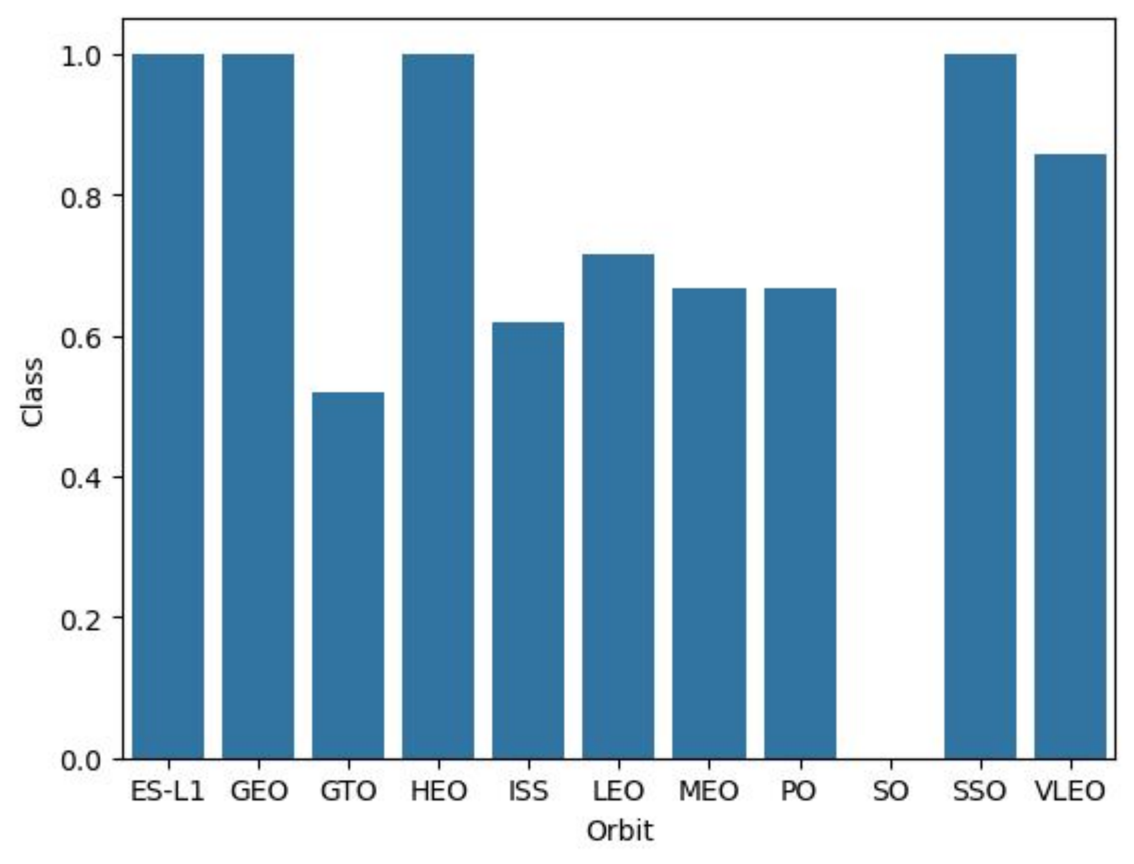
- Show a scatter plot of Payload vs. Launch Site
- Show the screenshot of the scatter plot with explanations



Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

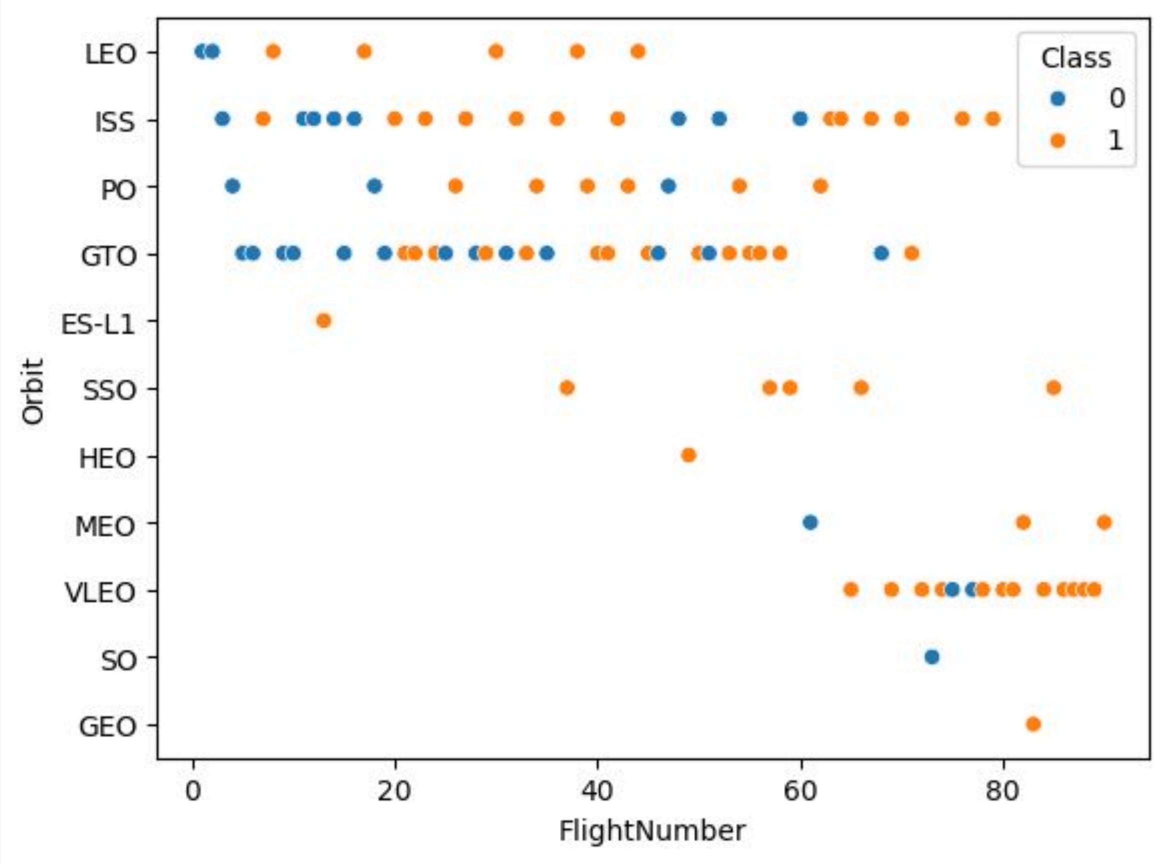
- Show the screenshot of the scatter plot with explanations



Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

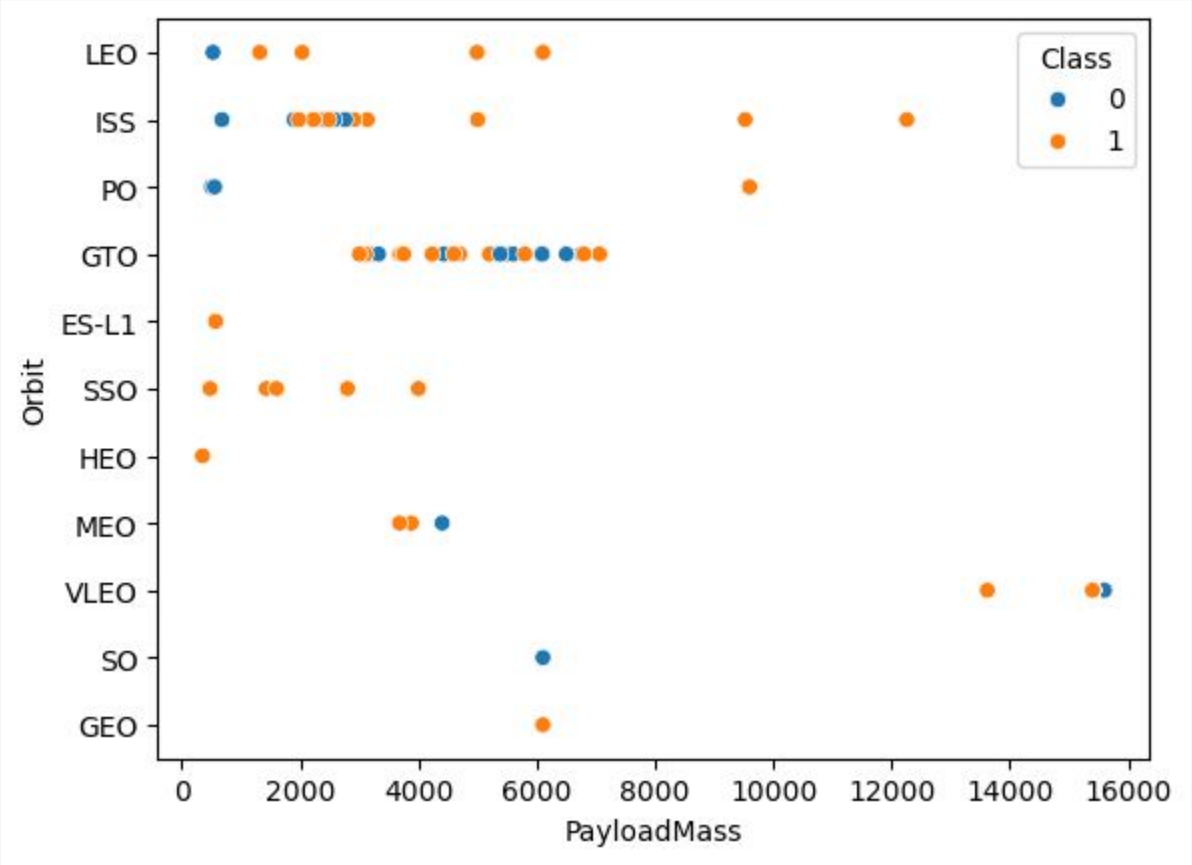
- Show the screenshot of the scatter plot with explanations



Payload vs. Orbit Type

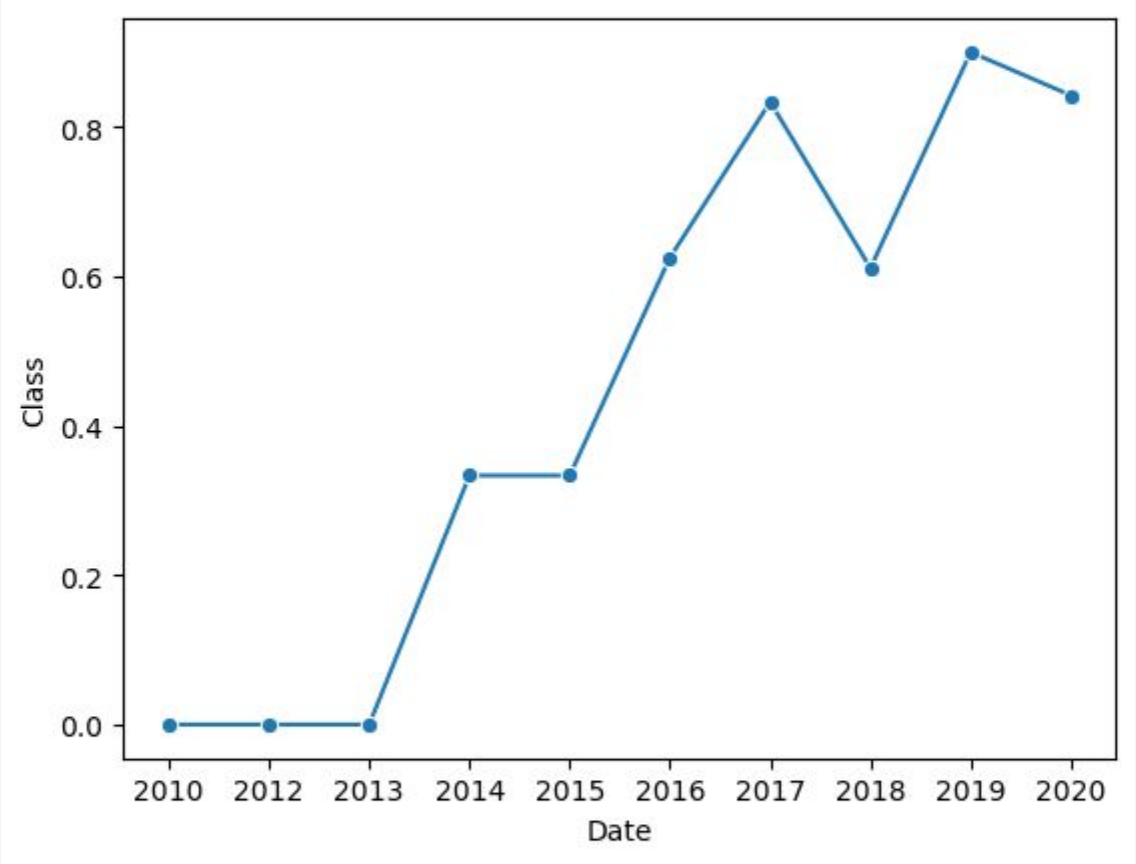
- Show a scatter point of payload vs. orbit type

- Show the screenshot of the scatter plot with explanations



Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Show the screenshot of the scatter plot with explanations



All Launch Site Names

The figure shows a list of all launch sites names

```
%sql select distinct Launch_site from SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Here are 5 records where launch sites begin with `CCA`

```
%sql select * from SPACEXTABLE where Launch_site like 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677

Total Payload Mass

the total payload carried by boosters from NASA

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
```

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

Dates of the first successful landing outcome on ground pad

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome like 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome like 'Success (drone ship)' and PAYLOAD_M
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(*) as total from spacextable group by Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Names of the booster which have carried the maximum payload mass

```
%sql select Booster_Version from spacextable where payload_mass__kg_ = (select max(payload_mass__kg_) from :
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select substr(Date,6,2) as month_name, Booster_version, Launch_Site, Landing_Outcome from spacextable w
```

```
 * sqlite:///my_data1.db
Done.
```

month_name	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Rount of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%sql select Landing_Outcome, count(*) AS OutcomeCount from spacextable where date between '2010-06-04' and
```

```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	OutcomeCount
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Map of Launch Sites

Launch Sites in US

<Folium Map Screenshot 2>

Replace <Folium map screenshot 2> title with an appropriate title

Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map

Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>

Replace <Folium map screenshot 3> title with an appropriate title

Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

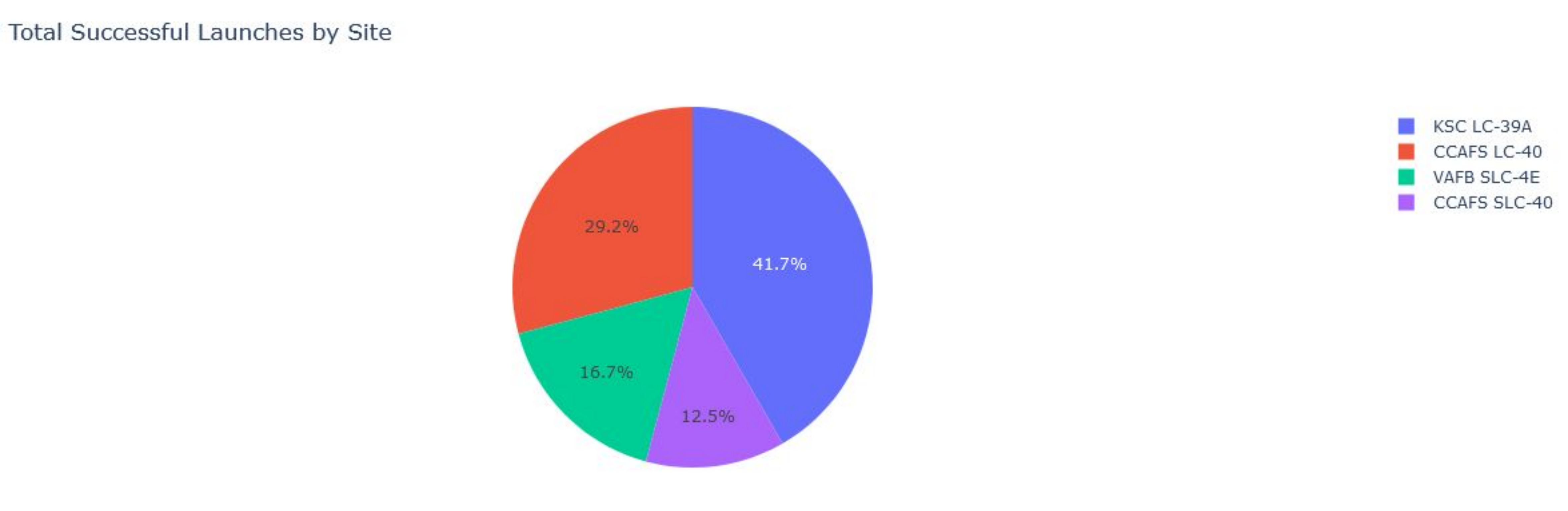
Explain the important elements and findings on the screenshot

Section 4

Build a Dashboard with Plotly Dash

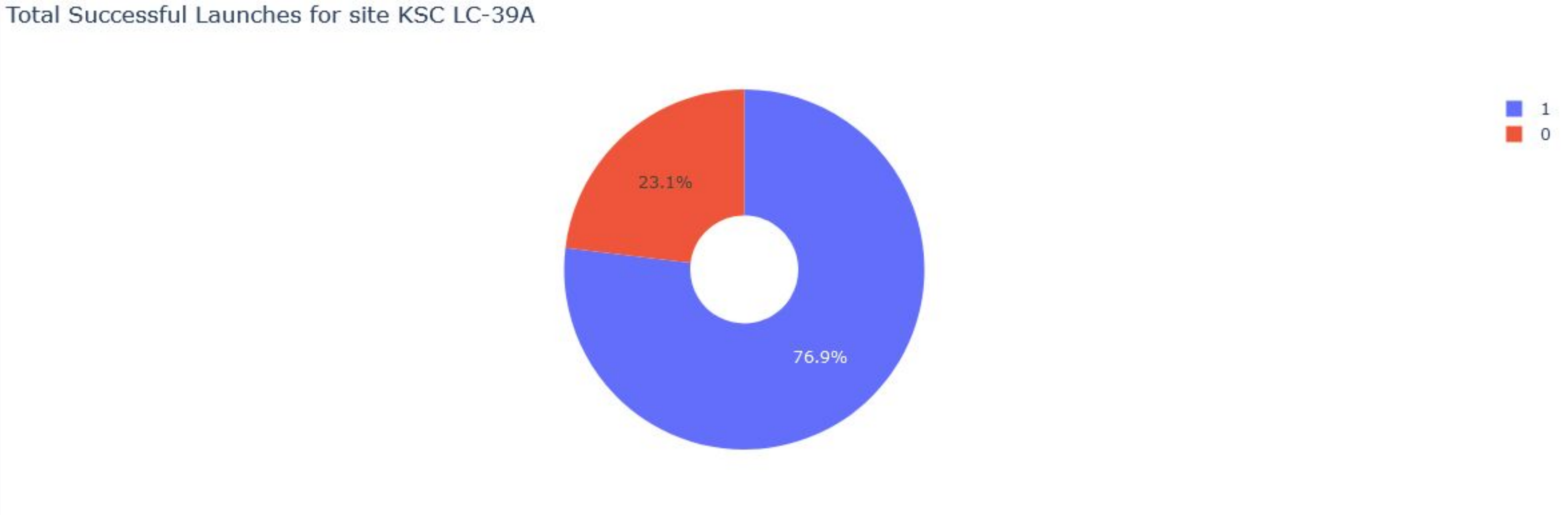
Launch success for all Sites

Shows the success proportion of all the sites



Site with highest success rate

Launch site with highest launch success ratio



Payload vs. Launch Outcome

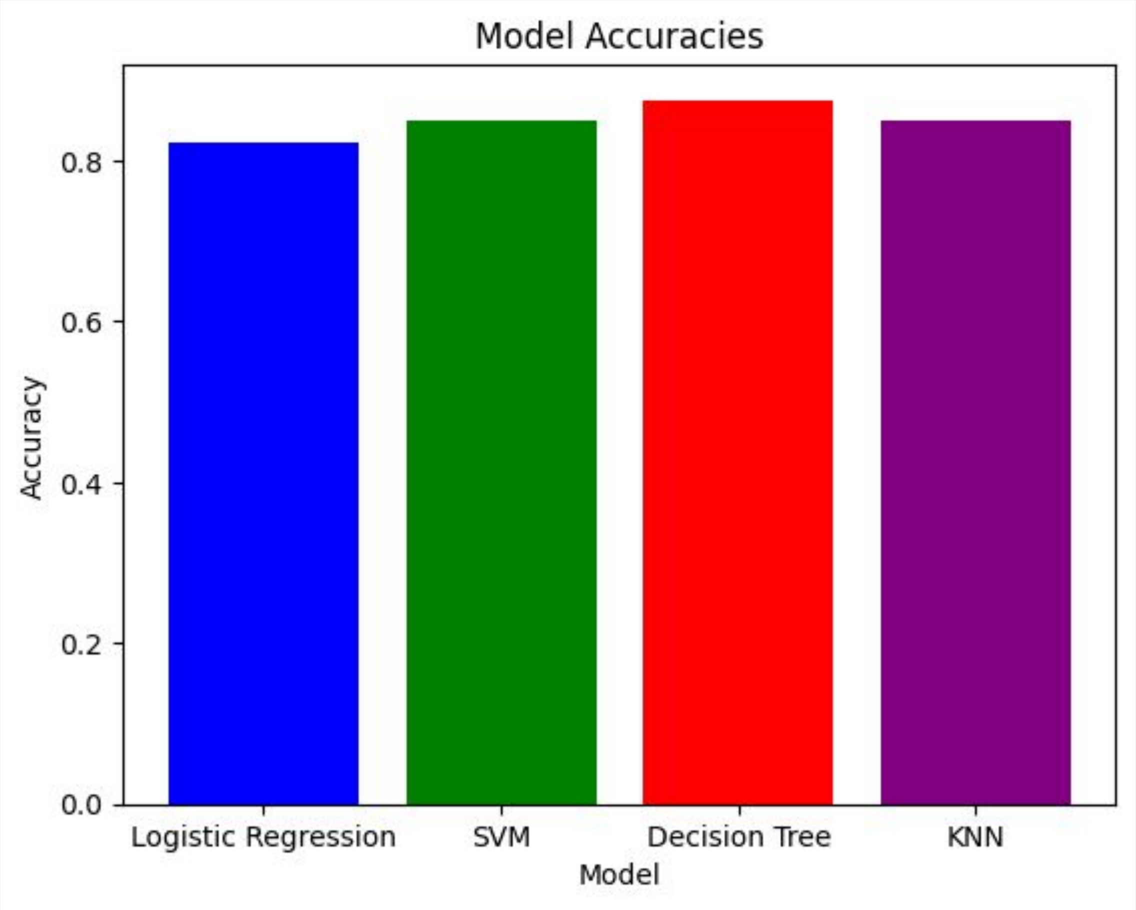
Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

Section 5

Predictive Analysis (Classification)

Classification Accuracy

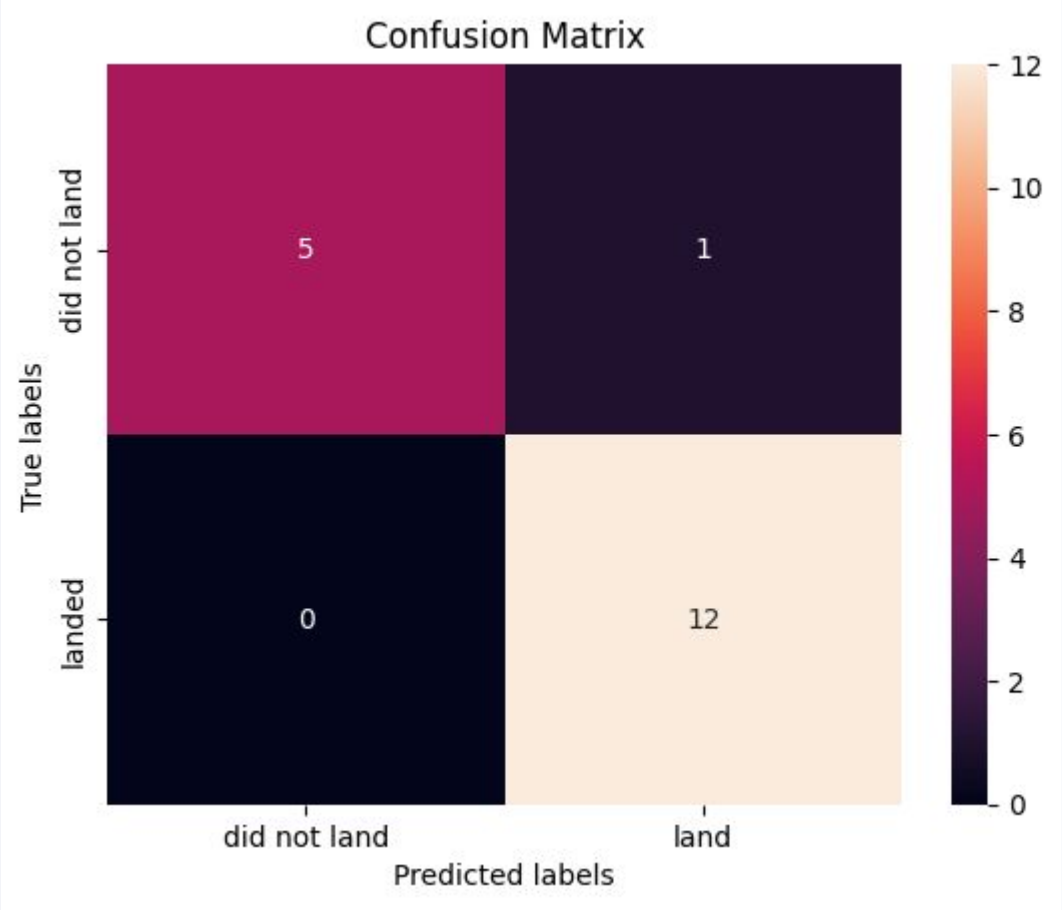
- We see that Decision Tree has the best accuracy for prediction model in this case



Confusion Matrix

- The confusion matrix for decision tree

We see it is less prone to false classifications than other models



Conclusions

- We were able to determine cost efficiency of launces based on various factors. Ksc launch site has greater success and so do GEO and HEO orbit launches
- We are able to see if Space X will reuse their first stages as they increase with time
- We can use decision tree model to determine success of first stage

…

Appendix

https://github.com/Barracuda108/Data-Science-Capstone.git

Thank you!